AD NUMBER
ADB345102
LIMITATION CHANGES
TO: **Approved for public release; distribution is unlimited.**
FROM: **Distribution: Further dissemination only as directed by Defense Advanced Research Projects Agency, Attn: DARPA/IPTO, 3701 N. Fairfax Dr., Arlington, VA 22203, SEP 2008, or higher DoD authority.**
AUTHORITY
DARPA, 11 Feb 2009

Portable Language-Independent Adaptive Translation from OCR

Quarterly R&D Status Report No. 4

Contractor:	**BBN Technologies** 10 Moulton Street, Cambridge, MA 02138
Principal Investigator:	Mr. Prem Natarajan Tel: 617-873-5472 Fax: 617-873-2473 Email: pnataraj@bbn.com
Reporting Period:	1 July 2008 – 30 September 2008

This material is based upon work supported by the Defense Advanced Research Projects Agency
DARPA/IPTO
Portable Language-Independent Adaptive Translation from OCR
MADCAT Program
ARPA Order No: X 103
Program Code: 7M30
Issues by DARPA/CMO under Contract #HR001-08-C-0004

EXECUTIVE SUMMARY

This is the fourth R&D quarterly progress report of the BBN-led team under DARPA's MADCAT program. The report is organized by technical task area.

1.1 Pre-Processing and Image Enhancement [BBN, Polar Rain, UMD, SUNY]

Page line detection and removal [BBN]: We implemented a ruled-line detection and removal algorithm which operates by first dividing an input image into 10 equal vertical stripes. Ruled-lines are detected at the stripe-level using the projection profile of the intensity. The detected ruled-lines are classified and marked as "black" or "white" using heuristics. If a non-ruled-line pixel is black and adjacent to a ruled-line pixel, then the pixels starting from the one immediately connected to the non-ruled-line pixel up to the one at the center of the ruled-line are marked as "black." The remaining pixels are marked as "white."

Page line removal and Shape-DNA restoration [Polar Rain]: This quarter, we configured our page line detection and removal software on MADCAT training images. We also optimized shape-DNA restoration on MADCAT data. The page-line removal and restoration software was provided to BBN for experimentation.

Other work during this quarter includes porting of SUNY's existing line-removal algorithm to work with Arabic handwritten documents, UMD's investigation of ruled-line removal for ANFAL and other challenging data types.

The impact of three line-removal algorithms on recognition performance is described in section 1.2.

1.2 Text Recognition [BBN, Argon, Columbia, SUNY]

Improvements in HMM based OCR [BBN]: In this quarter, we performed experiments on the training data released by LDC, and explored several techniques to reduce the word error rate (WER).

Updating Glyph Models with Additional Data: During the 4-month period June-September 2008, LDC released a total of 8253 scanned images of handwritten Arabic text of newswire articles, weblog posts, and newsgroup posts, along with the corresponding ground truth annotations. We measured the effect of incrementally increasing the amount of training data used for glyph modeling on text recognition performance. A separate set of 442 images released by LDC was split into two parts; one was used for development and the other for testing. Table 1 shows the performance on the test set with varying amounts of training data for glyph modeling. The WER was measured by detaching punctuations and sequences of digits from other words to which they may be attached.

Number of Training Images	Number of Training Authors	%WER	
		Authors in Training	Authors not in Training
848	10	51.3	36.3
3371	20	41.1	31.1
5288	38	41.6	29.4
8253	58	43.8	27.6

Table 1: OCR Performance with Glyph models trained on different amounts of training data

Unsupervised Adaptation: We performed MLLR adaptation of the glyph HMMs using the text recognition output of each page in the test set. The updated models were then used to re-decode the given page. As shown in Table 2, decoding with adapted models resulted in a relative improvement of 5%.

System	%WER
PACE Features	36.2
+ Unsupervised Adaptation	34.4
+ Gradient & Concavity Features	31.5

Table 2: Summary of Text Recognition Improvements on test set

Feature Extraction Improvements: In this quarter, we explored two new *structural* features for text recognition – Directional Element Features (DEF) and Gradient-Structure-Concavity (GSC) features. SUNY delivered their GSC feature extraction module which was integrated into BBN's text recognition

system. Since the baseline of a word image may fluctuate within portions of the same word, we algorithmically tightened the upper and lower boundaries of the sliding window used to compute the DEFs and GSC features. Used in combination with the baseline PACE (percentile, angle, correlation and energy) features, the DEFs provided a 4% relative improvement in WER while the best performance was seen with the combination of gradient and concavity features (8.4% relative improvement in performance).

Language Model (LM) Rescoring: We explored an initial version of word and word-part LM provided by Columbia. A modest gain in text recognition performance was observed by using a word part LM trained on the Arabic Gigaword corpus.

Slant Correction: In order to normalize the hand-written documents by different authors, we pre-processed the images by automatically correcting the slant in each word image by measuring the relative pixel organization along the perimeter of connected components and then using these statistics to re-organize each pixel position so as to reduce the overall slant in the image. The slant corrected images used for text recognition did not result in any improvements. We believe this is because the statistics used to perform slant correction on images are unreliable as they are computed only using the word images. We intend to revisit the slant normalization task again in the coming months.

Page-line Removal: We found that the performance of the text recognition system was considerably worse on pages with horizontal rules compared to pages without such rules. We pre-processed the images to remove lines using three line-removal algorithms from: a) Polar Rain, b) SSUNY, and c) BBN. The results in Table 3 on the test set pages containing ruled lines compare the three algorithms against a baseline without any line-removal. The SUNY and BBN algorithms result in a small improvement over the baseline, with SUNY's technique giving a relative improvement of 1.8%.

Line Removal Algorithm	% WER
None	38.1
SUNY	37.4
BBN	37.9
Polar Rain	39.3

Table 3: Comparison of performance of different line removal algorithms on test set

In Figure 1, we show an instance of line-removal performed by the three algorithms on a section of an image. The SUNY algorithm successfully removes lines from the image, but doesn't perform noise-removal as is done by the BBN and Polar Rain algorithms. The BBN and Polar Rain algorithms both remove some pixels associated with character glyphs which are connected to the line while performing line removal resulting in disconnected character segments in the image. Since the feature extraction algorithm does not rely on connected component analysis, the creation of disconnected components does not significantly affect system performance.

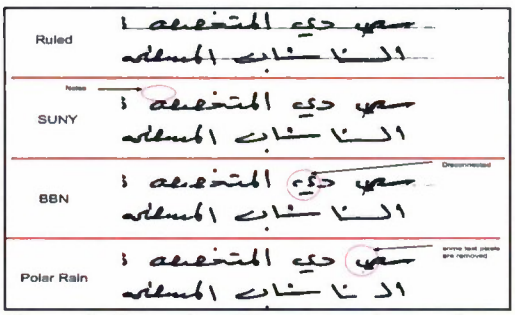

Figure 1: Examples of line removal using the different algorithms.

1.3 Logo Recognition [BAE]

We investigated the feasibility of two approaches for logo recognition. We developed a particular form of the trace transform which uses pixel value transitions to replace the line integrals of the radon transform. We established the rotational invariance of the transform, and investigated the effect of logo translation on transform appearance. We developed a translation-independent approach to trace transform matching and performed preliminary investigation into matching performance. We further developed an approach to estimate logo scale using this trace transform, and generated logo scale estimation results.

We also explored the in-house developed Alpha-Rooted Phase Correlation (ARPC) matching approach in the Fourier log-polar domain with the goal of translation, scale, and rotation-invariant logo recognition. We investigated the tolerance of Fourier log-polar domain matching as a function of logo rotation and determined that a combination of the interplay between digital sampling and rotation created artifacts that degraded recognition performance.

1.4 Integration with GALE Machine Translation [BBN]

The MT system used for Arabic to English translation is BBN's hierarchical MT system (HierDec) used by the AGILE team in GALE. Since the MADCAT data consists of a combination of Newswire and Web data, we ran experiments with four different MT systems tuned on either Newswire or Web genre data with and without discriminatively determined corpus weights. The system tuned on web genre with corpus weights out-performed the other three systems on the combined web and newswire test set. We also performed system combination with confusion networks similar to the GALE system. The primary difference between system combination on MADCAT and GALE is that the systems being combined in the GALE framework come from various sites that use different translation methodologies, whereas for MADCAT system combination, we simply used four differently tuned hierarchical MT systems trained on the same data. The system weights were tuned for TERBLEU on a combination of Newswire and Web genre documents. As shown in Table 4, the combined system results in significant improvements over the single-best system across both Newswire and Web genres.

System	Mixed-Case TER	
	Newswire	Web
Single Best System	50.1	56.5
Combined System	49.4	55.7

Table 4: Genre-wise comparison of single best system and combined MT system on error-free text of a development set.

1.5 Evaluation System [BBN]

We worked on the MADCAT Phase1 Evaluation held in September 2008. The glyph model used in the evaluation system was trained on a total of 8,253 images from 58 different authors. Position dependent tied mixture (PDTM) HMM models were trained for a total of 176 unique characters. A trigram language model trained on 90 million words of the GALE corpus in combination with a 92K dictionary was used for recognition. The *n*-best list from the recognizer was re-ranked using a combination of the glyph hmm scores, and a stronger language model score than in the recognizer. The weights for re-ranking were tuned on the development set. The 1-best hypothesis from the re-ranked *n*-best list was used to adapt the means of the HMM model via MLLR estimation. We trained two text recognition systems – one trained on the percentile, angle, correlation and energy features, referred to as PACE system, and the other trained on the PACE and gradient and concavity features, referred to as the GCPACE system.

The MT systems used for translating OCR output are described in Section 1.4. To introduce variability in the systems being combined to produce the final translation output, we mixed-and-matched the 2 text recognition systems (PACE and GCPACE) and 4 MT systems to produce four systems for combination. The performance of the 4 systems on a held-out portion of the Development set is shown in Table 5. The combined system provided a relative gain of 4.3% in mixed-case TER over the single-best system.

OCR System	MT System	%WER	Mixed-case TER	Mixed-case BLEU	METEOR
GCPACE	Web, Corpus Weights On	31.5	65.6	18.4	45.5
PACE	Web, Corpus Weights Off	34.4	67.0	17.7	44.0
PACE	Newswire, Corpus Weights On	34.4	67.3	17.1	43.8
GCPACE	Newswire, Corpus Weights Off	31.5	66.5	17.5	45.2
System Combination			62.8	19.5	46.0

Table 5: Phase1 Pilot Evaluation system results on text recognition output on the MADCAT test set selected from MADCAT Dev/test Part 1 release.